Review/Synthèse

Design-based and model-based inference in survey sampling: appreciating the difference

Timothy G. Gregoire

Abstract: Model-based ideas in finite-population sampling have received renewed discussion in recent years. Their relationship to the classical ideas in sampling theory do not appear to be universally well understood by samplers in applied disciplines such as forestry, and ecology more broadly. The two inferential paradigms are constrasted, and explanations are supplemented with examples of discrete as well as continuously distributed populations. The treatment of spatial structure is examined, also.

Résumé : Les idées sur l'échantillonnage des populations finies basées sur des modèles ont été le sujet de nouvelles discussions depuis quelques années. Leurs liens avec les idées classiques en théorie de l'échantillonnage ne semblent pas être universellement bien compris par les personnes effectuant de l'échantillonnage dans des disciplines appliquées telles que la foresterie et, plus généralement, l'écologie. Les deux paradigmes d'inférences sont comparés et les explications sont accompagnées d'exemples basés sur des populations distribuées de façon discrète ainsi que continue. Le traitement de la structure spatiale est aussi examiné.
[Traduit par la rédaction]

1. Introduction

The purpose of this article is to elaborate and contrast design-based with model-based inference in the context of survey sampling and the estimation of population values. It is aimed towards those who use sampling methods for purposes of research or public inquiry, who are uncertain of the essential differences between the two inferential paradigms and how those differences affect the interpretation of survey results. As Schreuder et al. (1993, p. 205) pointed out, the important issue is the recognition that both paradigms have a solid theoretical basis — they simply differ. It is hoped that by better understanding these differences, forestry and ecological scientists will appreciate the comparative advantages and disadvantages of using the sample design versus a postulated model as the basis for scientific inference, with the result that a more informed choice of one or the other can be made.

Since Neyman's (1934) famous paper read before the Royal Statistical Society, the literature on survey sampling, indeed the practice of same, has been dominated by an inferential paradigm wherein inference is independent from any assumptions about population structure and distribution because it relies instead on the distribution of all possible estimates permissible under the sampling design. The design is crucial for inference. The classic texts on sampling such as Hansen et al. (1953), Sukhatme and Sukhatme (1970), and Cochran (1977) expound this paradigm of design-based inference. Survey results from resource inventories in forestry and ecology have long been interpreted by the design-based perspective.

Alternative to the design-based framework is one in which a model serves as the basis for inference about population parameters in the context of survey sampling (cf. Brewer 1963; Royall 1970; and Cassel et al. 1977). While resource specialists in forestry and ecology are generally less familiar with model-based inference, there is a growing body of work within these disciplines that embraces this alternative inferential framework. Within forestry Mátern's (1960) classic work on spatial variation is probably the premier example of model-based inference. Rennolls (1981, 1982), Wood et al. (1985), Mandallaz (1991), Schreuder et al. (1993), Kangas (1994), and Eriksson (1995*b*) have also investigated the utility of model-based inference in forest inventory. These works notwithstanding, current literature in forestry and ecology indicate much ongoing confusion about the distinction between these two modes of inference. A failure to appreciate the underpinnings of one or the other mode of inference could lead to needless abandonment of a survey design that might otherwise be ideal for purposes of scientific inquiry. Of greater concern is the possibility that unwarranted or mistaken assumptions may lead to inferences whose validity does not withstand scientific scrutiny. The level of detail and technical complexity in this article is considerably less than that of Cassel et al. (1977), yet it is sufficiently explicit to highlight the differences, and where apt, the similarities of the two frameworks for inference. Illustrative scenarios of continuous and discrete populations are provided as examples. Extensive

Received March 19, 1998. Accepted October 5, 1998.

T.G. Gregoire. School of Forestry and Environmental Studies, Yale University, 360 Prospect Street, New Haven, CT 06511-2189, U.S.A. **e-mail:** timothy.gregoire@yale.edu

presubmission review of the material presented here indicated that the level of detail and illustrative derivations were trivially apparent and a distraction to some of those well versed in the relevant issues, while it was revealing and occasionally challenging to those that had never thoroughly considered the issues previously. The many numbered examples and remarks are intended to make the material understandable to a broad midrange of potential users. Suggestions for further reading to supplement selected issues raised in this article are made in the last section.

Remark 1

I conjecture that confusion about model-based inference has been fostered in part by an uncritical use of the terms “model-based sampling” and “model-based estimation” in contexts where model-based inference was intended. In my view, sample selection cannot be inherently model based. Sampling may be described aptly as being probabilistic, or sequential, or purposive, for instance, as these terms imply broadly the manner in which population elements are selected into the sample. One might describe the design of the sample as being model based, and it is perhaps this that Schreuder and Williams (1995) intended by the phrase “ . . . such model-based sample selection methods as purposive sampling should yield reliable estimates.”

Estimation, in contrast with sample selection but similar to the choice of sample design, may appeal to a model structure, as with the well-known ratio and regression estimators. But I question the utility of distinguishing between design-based and model-based estimators: properties of the linear regression estimator of the population total, say, can be deduced with respect to the design or with respect to a presumed model, which makes it rather equivocal to label the estimator itself as being design based or model based. The crucial distinction between design-based and model-based alternatives is whether inference, not the estimator, is based on the model. Hansen et al. (1983) proposed that estimation based on a model be termed model based and that inference based on a model be labeled model dependent. This semantic distinction has not been embraced widely in subsequent statistical literature, and it is not one to which I adhere. Specifically, in the remainder of this article, all comparisons of alternative inferential paradigms will be termed design based or model based as appropriate to the context.

2. Notation

In the discrete case, the population $\mathcal{U}$ consists of N elements. Associated with the kth of these elements is a value, Y_k. For multiresource surveys, Y_k may be a vector value, where each component of the vector is the value corresponding to a distinct resource. For the present, however, I concentrate upon the special case where Y_k is scalar. Associated with the kth element of the population is a $p \times 1$ vector, X_k', of auxiliary information. Even in the case where Y_k is scalar, X_k' may be vector valued, corresponding to multiple sources of auxiliary information.

The objective is to estimate some function, say $g(T_y)$, of the population total, where

$$T_y = \sum_{k\in\mathcal{U}} Y_k$$

In the simplest case $g(T_y)$ is the identity function, so that T_y is itself the target parameter. In other cases the target may be the population mean value

$$g(T_y) = \frac{T_y}{N}$$

or the target may be the population ratio

$$g(T_y) = \frac{T_y}{T_x} = R$$

Let I_k be an indicator of sample inclusion: $I_k = 1$ if the kth element is selected into the sample, and $I_k = 0$ otherwise. Let Ω symbolize the set of all possible samples permissible under the sampling design, and let $p(s)$ denote the probability of selecting sample S under this design. The probability of including the kth population element into a sample is by definition

$$[1] \qquad \pi_k = \text{Prob}(I_k = 1)$$

$$= \sum_{S\in\Omega} I_k p(s)$$

$$= \sum_{S\in\Omega_k} p(s)$$

where Ω_k is the subset of Ω comprising all those samples that include the kth population element. For example under simple random sampling (SRS) without replacement, Ω_k consists of $\frac{(N-1)!}{(n-1)!\,(N-n)!}$ equally likely samples, each with probability $p(s) = \frac{n!\,(N-n)!}{N!}$. Thus

$$\pi_k = \frac{(N-1)!}{(n-1)!\,(N-n)!} \times p(s) = \frac{n}{N}$$

When sampling with replacement, whether by SRS or not, $\pi_k = 1 - (1 - p_k)^n$, where p_k is the probability of selecting the kth element in each of the n draws. Notice that $\pi_k < np_k$ whenever $n > 1$, although for small p_k and large n, $\pi_k \approx np_k$.

The number of distinct elements in a sample is denoted by $\nu = \sum_{k\in\mathcal{U}} I_k$. When sampling with replacement for a fixed sample size of n elements, $\nu \le n$ whereas without-replacement sampling ensures that $\nu = n$. With-replacement designs usually result in ν being random while n is fixed whereas other designs result in n being random, e.g., Poisson sampling. In the former, $E[\nu] = \sum_{k\in\mathcal{U}} \pi_k$, while in the latter,

$E[n] = \sum_{k\in\mathcal{U}} \pi_k$, and neither expected value will necessarily be integer valued. Obviously when n is fixed a priori by design, as in SRS without replacement, $n = E[n] = \sum_{k\in\mathcal{U}} \pi_k$.

In the continuous case,

$$T_y = \int_{\mathcal{A}} Y(Z)\,\mathrm{d}Z$$

where Z specifies the spatial location within the region $\mathcal{A}$ over which the population is dispersed, and $Y(Z)$ indicates the spatial response surface at location Z. The areal extent of the population is thus

$$A = \int_{\mathcal{A}} \mathrm{d}Z$$

so that the population mean value is $\mu_y = T_y/A$. The (continuous) probability density of the sampling procedure is denoted by $f(Z)$. A uniform density is common, viz. $f(Z) = 1/A$; however, $f(Z)$ may be something other than the uniform density, as in applications of importance sampling (Gregoire et al. 1993, 1995) where it is proportional to a function of auxiliary information. Stevens (1997) provided a thoughtful and illuminating examination of variable probability sampling designs for continuous domains.

3. Design-based inference

In the design-based framework, the population is regarded as fixed whereas the sample is regarded as a realization of a stochastic process. In virtually all the standard texts on sampling, such as those cited in the introductory section, inference is based on the distribution of estimates generated by the sampling design and free of any assumptions about the distribution of Y_k (or $Y(Z)$ in the continuous case) values in the population (cf. Sărndal 1978; Rao 1997). This distribution is known as the randomization distribution; the derivation of this term is obscure but is likely attributable to Fisher (1935; cf. Box and Anderson 1955, p. 3). A useful concept is that of a *reference distribution* (Fisher 1956, p. 77; Rao 1985), as it permits an illuminating contrast with alternative bases of inference. In the context of design-based inference the distribution to which we refer in order to infer the statistical properties of estimators is the distribution of estimates that results from all possible samples permissible under the sampling design. For example, Stuart (1976) enumerated the 15 without-replacement samples of size $n = 2$ from a population of $N = 6$. The 15 estimates corresponding to these samples is the *reference set* and their distribution is the *reference distribution*.

Plausible models relating Y to X assist during the design stage to help craft an apt design (e.g., probability proportional to size, or pps, sampling) or an efficient estimator (e.g., a generalized ratio estimator of T_y), but inference remains firmly rooted in the design. With the model-assisted approach, as it has been called, one tries to devise estimators with good design-based properties and that nonetheless appeal to a descriptive model. Sărndal et al. (1992) is perhaps the premier reference to the model-assisted approach to survey sampling, and these authors' summary statement (p. 227) is authoritative:

> The role of the model ξ is to describe the finite population scatter. We hope the model ξ fits the population reasonably well. We think that the finite population looks as if it might have been generated in accordance with the model ξ. However, the assumption is never made that the population was really generated by the model ξ. Our conclusions about the finite population parameters are therefore independent of model assumptions.

In the design-based framework, the probabilistic nature of the sampling design is crucial, as it is the only source of randomness ascribed to each of the possible samples in Ω. This is not the case in the model-based approach to inference, as shown in the next section. Inasmuch as a design such as SRS does not necessarily imply the use of one particular estimator or another, the stipulation of both design and estimator is necessary in order to be unequivocal. The combination of a sampling design and estimator is known as a *sampling strategy* in current vernacular. When $E\left[n\right] = \nu = n$, an estimator, say $g(\widetilde{T}_y)$, is said to be consistent if $g(\widetilde{T}_y) = g(T_y)$ when $n = N$. Also known as Fisher consistency (Fisher 1973, p. 150), this concept is distinctly different from a limit in probability that one encounters in a text on mathematical statistics such as Bickel and Doksum (1977). Hájek (1981, p. 40) labeled this property "representativeness". Brewer (1994) pointed out that the notion of consistency enunciated by Neyman (1934) is "a very different concept" than that currently accepted.

Remark 2

One occasionally hears of "biased sampling," a term that is technically incorrect when applied to a probability sample wherein $\pi_k > 0$ for $k = 1, \ldots, N$. Bias is a property of an estimator, not a sample selection scheme; see Smith (1991) for an elaboration. Clearly, what this term implies is a sampling strategy that leads to biased results. The moniker "size-biased sampling" is well entrenched in the sampling literature, used loosely as a synonym for pps sampling. As Overton and Stehman (1995, p. 265) tactfully advised, this phrase " . . . must not be incorrectly interpreted to imply that sampling with probability proportional to size automatically creates bias since variable probability *samples* have no inherent bias. However, bias of the *estimator* is of concern . . . ."

Customarily the mean and variance of an estimator, i.e., the first two moments of the reference distribution, are of chief concern. Here, I derive both from first principles for the Horvitz–Thompson (HT) estimator, viz.:

$$[2] \qquad \widehat{T}_y = \sum_{k\in\mathcal{S}} \frac{y_k}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{y_k I_k}{\pi_k}$$

Because y_k is a fixed, not random, value in the design-based framework, the only thing in $\widehat{T}_y$ that is random is the inclusion of y_k into the sample, i.e., whether $I_k = 1$ or $I_k = 0$. Using $\widehat{T}_y(s)$ to signify the value of $\widehat{T}_y$ provided from a particular sample in Ω, one gets by definition

$$E\left[\widehat{T}_y\right] = \sum_{S\in\Omega} \widehat{T}_y(s) \times p(s)$$

$$= \sum_{S\in\Omega}\left(\sum_{k\in\mathcal{U}} \frac{y_k I_k}{\pi_k}\right) p(s)$$

$$= \sum_{k\in\mathcal{U}}\left(\sum_{S\in\Omega} \frac{y_k I_k}{\pi_k}\right) p(s)$$

$$= \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k}\left(\sum_{S\in\Omega} I_k p(s)\right)$$

$$= \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k}\left(\sum_{S\in\Omega_k} p(s)\right)$$

$$= \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k}\pi_k = T_y$$

where the defining relation for π_k in eq. 1 is used after the second to last line. I emphasize that this derivation of the expected value of $\widehat{T}_y$ is impervious to the distribution of the $\{y_k, k = 1, \ldots, N\}$ in $\mathcal{U}$ (henceforth denoted $y_k \in \mathcal{U}$).

The variance of $\widehat{T}_y$,

$$V\left[\widehat{T}_y\right] = \sum_{S\in\Omega}\left(\widehat{T}_y(s) - E\big[\widehat{T}_y\big]\right)^2 \times p(s)$$

$$= E\left[\widehat{T}_y^2\right] - \left(E\left[\widehat{T}_y\right]\right)^2$$

is derived in a parallel fashion. Because

$$\widehat{T}_y^2 = \left(\sum_{k\in\mathcal{U}} \frac{y_k I_k}{\pi_k}\right)^2 = \sum_{k\in\mathcal{U}} \frac{y_k^2 I_k^2}{\pi_k^2} + \sum_{k\neq k'}\sum_{\in\mathcal{U}} \frac{y_k y_{k'} I_k I_{k'}}{\pi_k \pi_{k'}}$$

and

$$\left(E\left[\widehat{T}_y\right]\right)^2 = T_y^2 = \sum_{k\in\mathcal{U}} y_k^2 + \sum_{k\neq k'}\sum_{\in\mathcal{U}} y_k y_{k'}$$

the following result is obtained:

$$[3] \qquad V\left[\widehat{T}_y\right] = \sum_{k\in\mathcal{U}} y_k^2\left(\frac{E\left[I_k\right]}{\pi_k^2} - 1\right)$$

$$+ \sum_{k\neq k'}\sum_{\in\mathcal{U}} y_k y_{k'}\left(\frac{E\left[I_k I_{k'}\right]}{\pi_k \pi_{k'}} - 1\right)$$

$$= \sum_{k\in\mathcal{U}} y_k^2\left(\frac{1-\pi_k}{\pi_k}\right)$$

$$+ \sum_{k\neq k'}\sum_{\in\mathcal{U}} y_k y_{k'}\left(\frac{\pi_{kk'} - \pi_k\pi_{k'}}{\pi_k\pi_{k'}}\right)$$

where $\pi_k = E\big[I_k\big] = E\big[I_k^2\big] = \mathrm{Prob}\big[I_k = 1\big]$, and where $\pi_{kk'} = \mathrm{Prob}\left[I_k = 1, I_{k'} = 1\right]$ is the joint inclusion probability of the kth and k'th elements together.

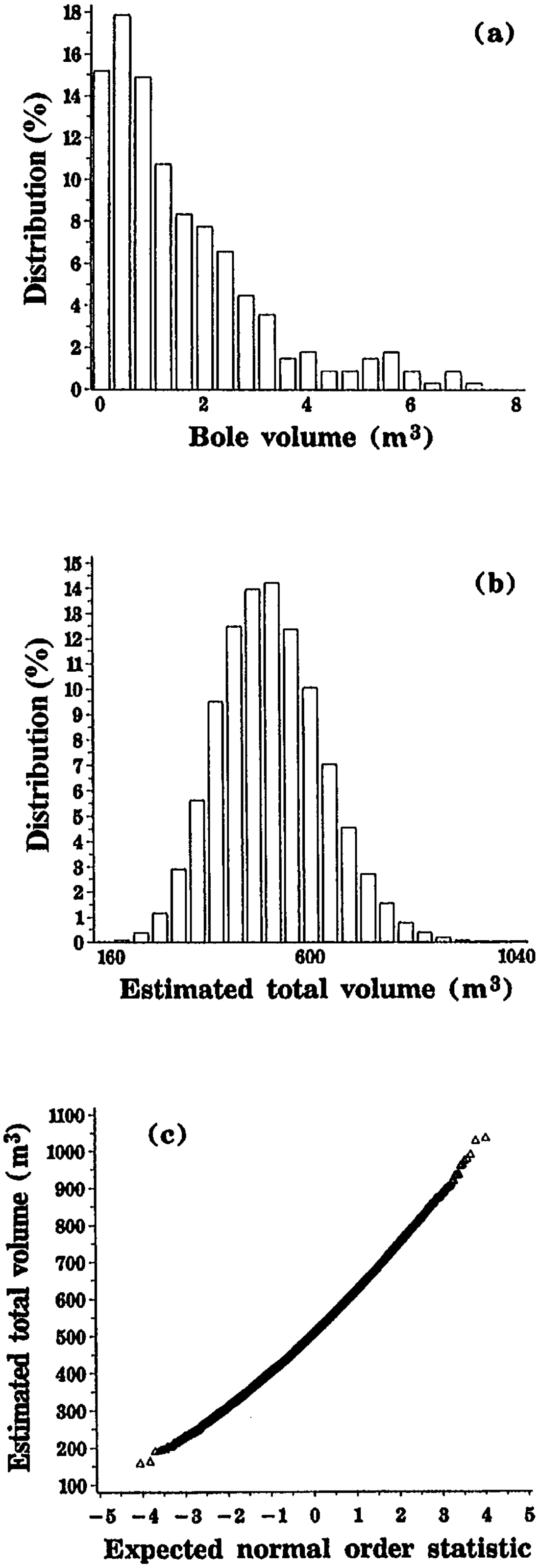
In the design-based framework the variance of $\widehat{T}_y$, and all other estimators, is the variance among the estimates from all possible samples in the reference set Ω — the reference distribution is the distribution of these estimates, not the distribution of the y_k values in the population. This is not meant to imply that the variance of an estimator of T_y will be unaffected by the variability of the $y_k \in \mathcal{U}$. Indeed for a few special cases the variance of an estimator can be expressed analytically as a function of the variance among the $y_k \in \mathcal{U}$, namely $\sigma_y^2 = N^{-1}\sum_{k\in\mathcal{U}}\left(y_k - \mu_y\right)^2$. But the essential point to remember is that even in the latter situation the variance of an estimator and the variance of the $y's$ are reckoned with respect to two different distributions in the design-based framework. Within this framework the variance of an estimator is not statistically dependent on the distribution of $y_k \in \mathcal{U}$. In the model-based framework, this is not the case.

For a spatially distributed discrete population such as trees in the forest, the spatial distribution of the population elements (trees) may affect the joint inclusion probabilities, but not necessarily. With a 3P/Poisson sampling design (cf. Schreuder et al. 1993), the joint inclusion probability is $\pi_{kk'} = \pi_k\pi_{k'}$, which demonstrates that it is impervious to the spatial distribution of trees. With a conventional fixed- or variable-radius plot sampling design, $\pi_{kk'}$ depends jointly on the horizontal distance between tree k and tree k', plot sizes, and plot shape. However, even in the latter case in which the spatial distribution of trees clearly affects their pairwise inclusion probabilities and thus will impact the variance of an estimator of T_y, the design-based variance of any estimator is unaffected by the presence, or lack thereof, of spatial covariance between y_k and $y_{k'}$. The notions of spatial distribution and spatial covariance, and their respective effects on design-based inference, must be kept distinct. As above, if the properties of an estimator are contingent upon, i.e., are derived in a manner which depends upon, a presumed spatial distribution of the $y_k \in \mathcal{U}$, then inference is no longer design based. I touch on this topic later in the section that deals with misconceptions that have been expressed about line intersect sampling.

Remark 3

For spatial covariance to impinge on inference, some structure for the population must be assumed, i.e., a population model, and this structure must be integrated into the inferential machinery. In the design-based setup, it is not, and

Fig. 1. (*a*) Frequency distribution of *Liriodendron tulipifera* L. trees, (*b*) distribution of $\widehat{T}_y$ from 25 000 samples of size $n = 20$, and (*c*) normal probability plot of $\widehat{T}_y$.



spatial correlation is an irrelevant issue. Indictments of classical sampling theory because " ... random sampling over spatial series does not insure that the sampling units are independent" (Bellehumeur et al. 1997) are simply wrong. Such statements fail to recognize that it is the independence of selections that is crucial in classical sampling theory and that establishes the validity of design-based inference. This misconception appears to be prevalent not only in the natural sciences, but also in the environmental and geological sciences (Brus and de Gruijter 1993).

Does $\widehat{T}_y$ follow a Gaussian distribution? Assuredly not, for even under SRS with replacement the sampling distribution is bounded from below by the estimate from the sample comprising n repeated selections of min($y_k \in \mathcal{U}$) and from above by the estimate resulting from the sample of n repeated selections of max($y_k \in \mathcal{U}$). Finite population versions of the Central Limit Theorem have been expounded, e.g., by Hájek (1960, as cited in Särndal et al. 1992, p. 59), which rely on carefully crafted asymptotic expansions. For reasons cited by Särndal et al. (1992), the reliance on asymptotic results has limited appeal when one seeks an interval estimator of a finite population parameter that attains its nominal coverage, or close to it. Nonetheless, it is customary to regard the erstwhile t statistic

$$\mathcal{T} = \frac{g\left(\widetilde{T}_y\right) - g(T_y)}{\sqrt{\widehat{v}\left[g\left(\widetilde{T}_y\right)\right]}}$$

as being approximately t-distributed using some appropriate estimator of variance $\widehat{v}\left[g\left(\widetilde{T}_y\right)\right]$. Lacking an exact theory for interval estimation, it is remarkable how closely confidence intervals based on $\mathcal{T}$ achieve their nominal coverage in many situations, as has been borne out empirically many times. Deviation of $\mathcal{T}$ from the t-distribution typically is most acute in the tail regions, so that nominal 80% intervals perform better than 90% intervals, which in turn perform better than 95% intervals, and so on, where performance is judged as the relative departure of the actual coverage rate from the nominal rate.

Skewed populations will affect the distribution of $\mathcal{T}$. Positive skew introduces a positive correlation between $g\left(\widetilde{T}_y\right)$ and $\widehat{v}\left[g\left(\widetilde{T}_y\right)\right]$, thus causing intervals based on the t-distribution to fail much more often from below than from above, where failing from below means that the upper endpoint of the interval is less than $g(T_y)$; negative skew induces the opposite effect. This phenomenon, mentioned by Royall and Cumberland (1985), was examined in detail by Gregoire and Schabenberger (1998) for the HT estimator $\widehat{T}_y$ and the ratio estimator $\widehat{T}_R = \left(\widehat{T}_y/\widehat{T}_x\right) T_x$ following SRS without replacement.

Table 1. Four-element population of Example 2.

k	1	2	3	4		
x_k	1.0	1.5	2.0	2.2	$T_x = 6.7$	$R = 1.013\,536$
y_k	1.001 79	1.639 06	1.757 26	2.392 58	$T_y = 6.790\,68$	$\rho_{xy} = 0.94$

Table 2. All possible samples and corresponding estimates for Example 2.

Sample	Sample x, y pairs	$\widehat{R}$	$\widehat{T}_R$	
a	1.0, 1.001 79			
	1.5, 1.639 06	1.0563	7.0775	
b	1.0, 1.001 79			
	2.0, 1.757 26	0.9197	6.1619	
c	1.0, 1.001 79			
	2.2, 2.392 58	1.0607	7.1070	
d	1.5, 1.639 06			
	2.0, 1.757 26	0.9704	6.5015	$E\left[\widehat{T}_R\right] = 6.7382$
e	1.5, 1.639 06			
	2.2, 2.392 58	1.0896	7.3005	$V\left[\widehat{T}_R\right] = 0.1584$
f	2.0, 1.757 26			
	2.2, 2.392 58	0.9881	6.6200	$\sqrt{\mathrm{MSE}\left[\widehat{T}_R\right]} = 0.3799$

Example 1

Figure $1a$ displays the frequency distribution of bole volumes from a set of 336 yellow-poplar trees (*Liriodendron tulipifera* L.). Typical of many biological populations, the positive skewness of this population is quite evident. From this population, 25 000 SRS samples of size $n = 20$ were selected without replacement using the method described by Bebbington (1975). The target parameter was total bole volume, which was $T_y = 516.9$ m^3. Figure $1b$ displays the empirical sampling distribution of the HT estimator, $\widehat{T}_y$, of T_y. The mean of this distribution differed from T_y by a mere -0.1%, and its variance differed from its analytically deducible value by less than 0.1%. The closeness of these results provides some assurance that the sampling error of the Monte Carlo experiment itself is negligible, despite the fact that the set of 25 000 samples used in the simulation is tiny compared with the set of more than 7×10^{31} samples in the sample space, Ω. The approximate normality evident in Fig. $1b$ is examined more specifically in the normal probability plot of Fig. $1c$. The concordance with normality is quite striking in this instance. Of the 25 000 confidence intervals constructed at a nominal 90% level, 88.9% actually included T_y, 8.1% missed from below, and 3.0% missed from above. ■

To recap some of the salient features of design-based inference: (*i*) the population is regarded as fixed in the sense that a fixed value, y, is associated with each element of the population (there is no notion that the population has been "randomized" by some agent or process), (*ii*) the reference distribution is the distribution of estimates generated by the combination of the design and a specific estimator, and (*iii*) the statistical properties of estimators are deduced from the probability weighted moments of the reference distribution.

Example 2

For sake of demonstration, consider a population of size $N = 4$ from which an SRS of size $n = 2$ is drawn without replacement for the purpose of estimating the population total, T_y. The paired x, y values are displayed in Table 1, along with the the population totals T_x, T_y, the population ratio $R = T_y/T_x$, and the correlation between x and y, ρ_{xy}. The sample space Ω comprises the six samples enumerated in Table 2; each sample is accompanied by the sample estimate $\widehat{R} = \widehat{T}_y/\widehat{T}_x$ and the ratio estimate $\widehat{T}_R = \widehat{R} \times T_x$ corresponding to each sample. Averaging over these six equally likely sample estimates yields $E\left[\widehat{T}_R\right] = 6.7382$, from which the design bias is computed to be a slight -0.7% of the targeted value $T_y = 6.7907$. In the design-based framework, $\widehat{T}_R$ is a biased estimator of T_y, irrespective of any stochastic process that actually may have generated the population. This is commonly misunderstood, perhaps after reading Cochran's (1977) statements of conditions under which the ratio estimator is a best linear unbiased estimator (BLUE). Yet Cochran also unequivocally asserted that BLUEness obtains only when one bases inference on the model $Y = \beta X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 X)$. The import of this assertion appears not to be appreciated by many who believe that design-based unbiasedness obtains under the same conditions that ensure model unbiasedness. This example dispels that notion, as the data analyzed here actually were generated by a $Y = \beta X + \epsilon$ model. I visit this topic again in Example 6*b* where this population is reanalyzed from a model-based perspective. ■

Example 3

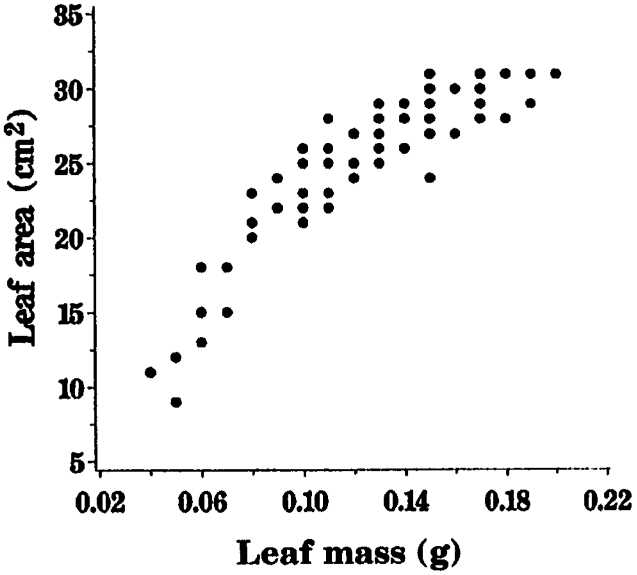
With a systematic sampling design, there are many ways in which auxiliary information, $x_k \in \mathcal{U}$, can be utilized in order to increase precision of estimation over that possible when the auxiliary information is ignored. Aside from the ratio estimator, an effective method is to sort the population in order of increasing value of the $x_k \in \mathcal{U}$ prior to sampling (cf. Sărndal et al. 1992). As with other uses of auxiliary information, the gain in precision to be realized increases the more strongly the auxiliary information, x, is correlated with the characteristic of interest, y. As long as the correlation is positive, then the act of ordering the population induces a correlation between adjacent values of y_k in the ordered population that is akin to a spatial correlation. Consider the population of $n = 64$ leaves shown in Fig. 2; the correlation between leaf area (y) and leaf weight (x) is $\rho = 0.89$, and the population totals are $T_y = 1582$ cm^2 and $T_x = 7.91$ g. Prior to ordering by increasing leaf weight, the correlation between adjacent leaf area values was 0.143; after ordering, it was 0.85. Precisely because this lag 1 correlation is so strong, the effect of ordering on estimation is to decrease the variation among samples, and hence to reduce the variance of an estimator of T_y. To be specific, consider the sampling strategy consisting of a 1-in-a systematic sampling design (each ath element is selected) coupled with the HT estimator of T_y. The 1-in-a design ensures that Ω consists of exactly a samples, and that $E[\nu] = E[n] = N/a$.[1] Because each element of the population can appear in but one of the possible samples, $\pi_k = 1/a, k = 1, \ldots, N$, and the $\widehat{T}_y$ estimator in eq. 2 collapses to

$$[4] \qquad \widehat{T}_y = a t_s$$

where $t_s = \sum_{k \in \mathcal{S}} y_k$. The variance of $\widehat{T}_y$ as shown in eq. 3 collapses to

$$[5] \qquad V\left[\widehat{T}_y\right] = a \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2$$

Fig. 2. Relationship between area and weight for a population of $N = 64$ leaves.



where $\bar{t} = T_y/a$. The last expression makes it quite evident that the variance of the estimator is the variance among estimates. However, the effect of ordering can be appreciated better by reexpressing eq. 5 in terms of the average correlation among pairs of y_k in the same systematic sample, ρ_w:

$$[6] \qquad V\left[\widehat{T}_y\right] = \frac{N^2 \sigma_y^2}{n} \left[1 + (n-1)\rho_w\right]$$

(see Cochran 1977, p. 209; Sărndal et al. 1992, p. 79). Ordering by size of the auxiliary variable is beneficial when it makes ρ_w negative and, it is hoped, lessens its value from what it is in the population's original order.

For this example, all 10 of the samples possible from a 1-in-10 systematic sampling from the original, haphazardly ordered leaf population were selected. Then the population was arranged in order of increasing leaf weight, and all 10 possible samples under this ordering were selected. From the unordered population, $\sqrt{V\left[\widehat{T}_y\right]} = 175.6$ cm^2 whereas from the ordered population, $\sqrt{V\left[\widehat{T}_y\right]} = 130.44$ cm^2 — a 26% increase in precision simply by ordering. The root mean square error of the ratio estimator, $\widehat{T}_R = \left(\widehat{T}_y/\widehat{T}_x\right) T_x$, with the ordered population is smaller yet at 53.1 cm^2. Clearly, ordering has an effect on precision, yet precision, or accuracy in the case of $\widehat{T}_R$, is reckoned solely on the basis of between-sample variation of the estimates, not on within-sample variation and covariation. ■

Despite being the dominant paradigm in survey sampling for well over half a century, inference on the basis of the sampling design has not met with universal acclaim. Hájek (1981, p. 34) complained "What relevance have samples that could

[1] The sample size, n, will be random unless $\text{mod}(N, a) = 0$.

have been drawn and their probabilities, if we know that they have not been drawn? Should not inference be based just on the particular sample that has been drawn?" Särndal (1978) speculated that "Someone like R.A. Fisher would more likely have considered that the randomization is important before, but not after, the data have been collected." Some statisticians view the randomness imposed by the sampler to be an artificial basis for inference (e.g., Basu 1978; Fisher 1956, p. 99). Sampling with probability proportional to size and the consequent assignation of differential weights to observations depending on their size is a tactic that, if not anathema, is regarded as at least peculiar by many, although it seems utterly natural to proponents of the design-based approach. (The notions of conditioning estimation on the observed sample and the disregard of sample probabilities, at least for the sake of inference, are features of the model-based approach.) Finally, when studying populations that change both in time and in composition, perhaps for the purpose of estimating growth or change, then the credibility of treating the population as fixed strikes some, e.g., Eriksson (1995*a*), as rather Procrustean.

4. Model-based inference

The fundamental difference between the design-based and the model-based approach to inference in survey sampling is that the values $y_1, \ldots, y_N$ are regarded as realizations of random variables $Y_1, \ldots, Y_N$ (Särndal 1978; Thompson 1997), and hence the population is a realization of a random process, called generically a model, a "superpopulation" model, or just a superpopulation. Not only may inference be concerned with one or more parameters of the survey population, e.g., $g(T_y)$ as in the preceding section, but also parameters, say θ, of the superpopulation. Because the presumption of a model broadens the inference space to include superpopulation parameters, it requires more assumptions than the design-based approach. But in this regard, it accords with nearly everything else one does in statistical estimation and prediction: a model is assumed based on prior experience and subject matter knowledge, the model is fitted to sample data according to some criterion (least squares, maximum likelihood, minimax risk), the goodness of fit is checked, alterations are made if deemed warranted, and eventually the results of the fitted model are proclaimed. This familiarity with fitting models to data is for some the appeal of model-based inference in the context of sample surveys (cf. Thomson 1978).

I view a model in a broad sense "to mean any assumptions about the structure of the population" (Smith 1994). The specification of the model may be quite exacting, e.g., the conditional distribution of Y on X, or it may be something comparatively unstructured, e.g., an assumption of a Poisson distribution of Y. Irrespective of the level of detail stipulated by the survey analyst, inference in the model-based approach stems from the model, not from the sampling design. The reference distribution is the distribution of Y for a given sample, not the distribution of Y over all possible samples. For example, estimator variance is reckoned conditionally upon the sample actually observed; it is the variability of the possible realizations of Y for the set of X values observed in the sample, where "possible realizations" are governed by the distribution of Y stipulated in the model. For valid inference in the model-based approach, sample selection still must be uninformative (see Särndal 1978) with respect to Y, but it indeed may be informative, even purposive, with respect to X and, depending on context, spatial location (Z).

In the design-based approach, no uncertainty about $g(T_y)$ remains when the entire population is censused; in the model-based approach, $g(T_y)$ is a random variable, and therefore, uncertainty about its distribution remains even after a population census, as the superpopulation parameters, θ, generally will still not be known with certainty.

Example 4

Consider the mean model given by

$$[7] \qquad Y_k = \mu + \sigma\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0,1), \quad \mathrm{cov}(\epsilon_k, \epsilon_{k'}) = 0$$

Under this model, $T_y \sim \mathcal{N}(N\mu, N\sigma^2)$. If a SRS of size n is selected, the estimator $\widehat{T}_y = N\bar{y}$ has a $\mathcal{N}(N\mu, N^2\sigma^2/n)$ distribution. Compare this result with the moments of $\widehat{T}_y$ under a design-based approach in which $E\left[\widehat{T}_y\right] = T_y$, a fixed constant, and $V\left[\widehat{T}_y\right] = N^2\sigma_y^2\left(\frac{1}{n} - \frac{1}{N}\right)$ when SRS is without replacement and $V\left[\widehat{T}_y\right] = N^2\sigma_y^2/n$ when SRS is with replacement. The differences are subtle at first glance, but meaningful from an inferential standpoint: in the design-based arena, T_y, $\mu_y = T_y/N$, and $\sigma_y^2 = N^{-1}\sum_{k\in\mathcal{U}}(y_k - \mu_y)^2$ are all population parameters whereas μ and σ^2 are model parameters in the model-based approach and T_y is a random variable. ■

Remark 4

For large N, one would expect that the mean of the realized population, μ_y, would be quite close in value to the mean, μ, of the stochastic process having generated the population, and likewise that the population variance, σ_y^2, would be close to the superpopulation variance, σ^2. The validity of inference in the model-based approach depends on how well Y_k accords with the stipulated model whereas no model of population behavior is presumed in the design-based approach.

Remark 5

Had the sample in Example 4 been selected purposively — say by choosing the $n/2$ elements at each end of an ordered (by something other than Y) list — the reference distribution under model [7] remains unaffected whereas the reference distribution under the design collapses to the single obtained value $\widehat{T}_y$. Model unbiasedness is unaffected by sample selection, yet design unbiasedness is not. For sake of inference under model [7], sample selection must be noninformative, in the sense as explained by Särndal (1978, p. 33), but it need not be probabilistic. The design is irrelevant to inference when the reference distribution is established by the model, yet obviously, it is all important to inference when the reference distribution is established by the design.

Example 5*a*

Consider the regression model with correlated observations given by

$$Y_k = \beta_0 + \beta_1 X_k + \sigma\epsilon_k,\ \epsilon_k \sim \mathcal{N}(0,1),\ \mathrm{cov}(\epsilon_k, \epsilon_{k'}) = \rho$$

Thus, $T_y \sim \mathcal{N}\big(\beta_0 + \beta_1 T_x, N\sigma^2[1 + (N-1)\rho]\big)$.

Example 5*b*

Consider the regression-through-the-origin model

$$Y_k = \beta X_k + \sigma\epsilon_k\sqrt{X_k},\ \epsilon_k \sim \mathcal{N}(0,1),\ \mathrm{cov}(\epsilon_k, \epsilon_{k'}) = 0$$

A weighted least squares fit of this model to the entire population, namely $Y_k, X_k, k = 1, \ldots, N$, yields

$$\widehat{\beta} = T_y/T_x = R$$

Using the postulated distribution for ϵ_k, it is evident that $\widehat{\beta}$ is unbiased for β:

$$E_{\mathrm{m}}\big[\widehat{\beta}\,\big] = \frac{E_{\mathrm{m}}\left[T_Y\right]}{T_X} = \frac{1}{T_X}\,E_{\mathrm{m}}\left[\sum_{k\in\mathcal{U}} Y_k\right]$$

$$= \frac{1}{T_X}\,E_{\mathrm{m}}\left[\sum_{k\in\mathcal{U}} \beta X_k + \sigma\epsilon_k\sqrt{X_k}\right]$$

$$= \frac{1}{T_X}\sum_{k\in\mathcal{U}} E_{\mathrm{m}}\left[\beta X_k + \sigma\epsilon_k\sqrt{X_k}\right]$$

$$= \frac{1}{T_X}\sum_{k\in\mathcal{U}} \left(\beta X_k + E_{\mathrm{m}}\left[\sigma\epsilon_k\sqrt{X_k}\right]\right)$$

$$= \frac{1}{T_X}\sum_{k\in\mathcal{U}} \left(\beta X_k + \sigma E_{\mathrm{m}}\left[\,\epsilon_k\,\right]\sqrt{X_k}\right)$$

$$= \frac{1}{T_X}\sum_{k\in\mathcal{U}} \beta X_k = \beta$$

where the subscript "m" indicates expectation with respect to the model. Since the model stipulates that $V_{\mathrm{m}}\left[\,Y_k \mid X_k\,\right] = \sigma^2 X_k$, then

$$V_{\mathrm{m}}\left[\,\widehat{\beta}\,\right] = \frac{1}{T_X^2}\sum_{k\in\mathcal{U}} \sigma^2 X_k = \frac{\sigma^2}{T_X}$$

Thus, $\widehat{\beta}$ follows a $\mathcal{N}(\beta, \sigma^2/T_X)$ distribution. ■

Two things highlighted by this example deserve repeated emphasis: (*i*) even after fitting the model to the entire population, the model parameters, β and σ^2, remain unknown; (*ii*) since properties of estimators are derived conditionally on the observed sample, X_k is treated as a constant throughout these derivations. In the design-based framework, an estimator of $g(T_y)$ is unbiased if its expected value with respect to the design coincides with $g(T_y)$. Obviously, this loses force in the model-based framework because $g(T_y)$ is itself a random variable, not a population parameter. In this framework, an estimator of $g(T_y)$ is said to be model unbiased when $E_{\mathrm{m}}\left[g(\widetilde{T}_y) - g(T_y)\right] = 0$.

Remark 6

How is one to interpret a statement such as "It should be mentioned that $\widehat{R}$ is an unbiased estimator of R if and only if the relationship between x and y is linear and goes through the $(0,0)$ point on the x–y axes?" (Shiver and Borders 1996, p. 160). Clearly, this statement is true when inference is based on a model that correctly specifies this relationship and when the distribution posited by the model is used as the reference distribution from which the properties of estimators are derived. Just as clearly, it is false when the randomization distribution is used as the basis for inference, irrespective of whether the model holds, e.g., see Example 2. Without further qualification, such a statement is at best ambiguous and at worst misleading. Survey designers who rely on the design-based context of Shiver and Borders, for example, to assert the design unbiasedness of the ratio estimator will discover such an assertion to be justly challenged.

Example 6*a*

The $N = 4$ population of Example 2 was generated by the model of Example 5*b* with $\beta = 1$ and $\sigma^2 = 0.05$. The BLUE estimator of β is $\widehat{R} = \bar{Y}/\bar{X}$, as first noted by Brewer (1963). For samples of size n:

$$V_{\mathrm{m}}(\widehat{R}) = \left(\frac{1}{\sum_{k\in\mathcal{S}} X_k}\right)^2 \sum_{k\in\mathcal{S}} V_{\mathrm{m}}\left(Y_k \mid X_k\right)$$

$$= \frac{\sigma^2}{\sum_{k\in\mathcal{S}} X_k}$$

and where the subscript "m" again indicates expectation with respect to the model. For maximum precision, therefore, one should choose that sample that maximizes $\sum_{k\in\mathcal{S}} X_k$. Such purposive selection is a foreign notion to design-based devotees, who will rely on stratification to achieve a similar result.

Example 6*b*

Under the model of Example 5*b*, the best linear unbiased predictor of T_y is the familiar $\widehat{T}_R = \widehat{R} \times T_x$. The variance

Table 3. Variance of $\widehat{T}_R$ under the model considered in Example 6*b*.

Sample	$\widehat{T}_R$	$V_{\mathrm{m}}[\widehat{T}_R - T_y]$
a	7.0775	0.5626
b	6.1619	0.4132
c	7.1070	0.3664
d	6.5015	0.3063
e	7.3005	0.2716
f	6.6200	0.1994

of $\widehat{T}_R$ under the model is defined to be

$$\begin{aligned} V_{\mathrm{m}}\left[\widehat{T}_R - T_y\right] &= V_{\mathrm{m}}\left[\left(\sum_{k\in\mathcal{S}} Y_k + \widehat{R}\sum_{k\in\mathcal{R}} X_k\right)\right. \\ &\qquad \left. - \left(\sum_{k\in\mathcal{S}} Y_k + \sum_{k\in\mathcal{R}} Y_k\right)\right] \\ &= V_{\mathrm{m}}\left[\widehat{R}\sum_{k\in\mathcal{R}} X_k - \sum_{k\in\mathcal{R}} Y_k\right] \\ &= \sigma^2\left(\sum_{k\in\mathcal{R}} X_k \Big/ \sum_{k\in\mathcal{S}} X_k\right)^2 \sum_{k\in\mathcal{S}} X_k \\ &\qquad + \quad \sigma^2 \sum_{k\in\mathcal{R}} X_k \\ &= \sigma^2 T_x \left(\sum_{k\in\mathcal{R}} X_k \Big/ \sum_{k\in\mathcal{S}} X_k\right) \end{aligned}$$

where $\mathcal{R}$ indicates the set of elements in $\mathcal{U}$ but not in $\mathcal{S}$. The variance of $\widehat{T}_R$ under the model for each of the six samples of Example 2 is displayed in Table 3. The model can be used to predict values of Y_k as $\widehat{Y}_k = X_k\widehat{R}$. These predicted values provide an unbiased estimator of σ^2, namely:

$$\widehat{\sigma}^{\,2} = (n-1)^{-1}\sum_{k\in\mathcal{S}}(Y_k - \widehat{Y}_k)^2/X_k$$

which can be used to compute an estimator of $V_{\mathrm{m}}\left[\widehat{T}_R - T_y\right]$:

$$\widehat{v}\left[\widehat{T}_R - T_y\right] = \widehat{\sigma}^2 T_x \left(\sum_{k\in\mathcal{R}} X_k \Big/ \sum_{k\in\mathcal{S}} X_k\right)$$

which Royall (1971) called “a useful, if crude, indicator of the after sampling uncertainty” in $\widehat{T}_R$.

Remark 7

Within a modeling framework, it is customary to determine optimal estimators, such as the BLUE in the preceding example. Doing similarly in a design-based framework is fruitless, as first established by Godambe (1955; also see Lanke 1975). ■

Example 7

Consider the simple mean model of Example 4 again. Suppose now that the SRS design is replaced by one in which the inclusion probabilities are unequal. The bias, where expectation is reckoned with respect to the model, of the HT estimator is

$$\begin{aligned} E_{\mathrm{m}}\left[\widehat{T}_y - T_y\right] &= E_{\mathrm{m}}\left[\sum_{k\in\mathcal{S}} \pi_k^{-1} Y_k\right] - T_y \\ &= E_{\mathrm{m}}\left[\sum_{k\in\mathcal{S}} \pi_k^{-1}\left(\mu + \sigma\epsilon_k\right)\right] - T_y \\ &= \sum_{k\in\mathcal{S}} \pi_k^{-1}\left(\mu + E_{\mathrm{m}}\left[\sigma\epsilon_k\right]\right) - N\mu \\ &= \mu\left(\sum_{k\in\mathcal{S}} \pi_k^{-1} - N\right) \end{aligned}$$

Thus, whenever $\sum_{k\in\mathcal{S}} \pi_k^{-1} \neq N$, the HT estimator is a biased estimator of T_y when inference is based on model [7] of Example 4. The bias may be slight in large samples: because $\sum_{k\in\mathcal{S}} \pi_k^{-1}$ is the HT estimator of N, it is design unbiased as an estimator of N, and consequently, its deviation from N is expected to be small in large samples. Nonetheless, the model bias of the HT estimator, $\widehat{T}_y$, is nonzero, in general, under unequal probability sample designs, a result that is anathema to many proponents of design-based inference. ■

A further contrast of the design-based approach and the model-based approach but with a continuously distributed population is considered next.

Example 8

The effect of harvesting and site preparation on soil temperature is an important concern of forest management. Let $\mathcal{A}$ denote a region recently harvested whose area is

$$A = \int_{\mathcal{A}} \mathrm{d}Z$$

where Z indicates location in the two-dimensional plane of $\mathcal{A}$. In the following, I abide by the convention used in the preceding section but not mentioned explicitly, of denoting a random variable in uppercase and a fixed quantity in lowercase. Suppose that the objective is to estimate the average soil temperature within $\mathcal{A}$. In the design-based framework the target parameter is $\mu_y = A^{-1}\int_{\mathcal{A}} y(z)\,\mathrm{d}z$, where $y(z)$ represents the soil temperature at location z. One can choose n locations independently and uniformly at random over $\mathcal{A}$, i.e., an SRS of n pairs of coordinates along orthogonal axes, in which case, each sample location Z is random (under the design), and then estimate μ_y by

$$\bar{y} = n^{-1}\sum_{k=1}^{n} y(Z_k)$$

where Z_k is the kth randomly selected location. The variance of $\bar{y}$ is

$$V\left[\,\bar{y}\,\right] = A^{-1}\int_{\mathcal{A}}\bigl(y(Z) - \mu_y\bigr)^2\,\mathrm{d}z$$

which is estimated unbiasedly by

$$\widehat{v}\left[\,\bar{y}\,\right] = \bigl(n(n-1)\bigr)^{-1}\sum_{k\in S}\bigl(y(Z_k) - \bar{y}\bigr)^2$$

As with Example 3, the spatial correlation is an irrelevant concern when variance is reckoned with respect to the randomization distribution.

In a model-based framework, one could envision a simple mean model

$$Y(z) = \mu + \epsilon, \qquad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\mathrm{cov}\bigl(Y(z), Y(z+h)\bigr) = \sigma^2 C(h), \qquad h > 0$$

In this setup, $C(h)$ denotes the spatial correlation in soil temperatures separated by a distance h on the ground. Interest may focus on the fixed parameter, μ, or the finite population value

$$\bar{Y} = A^{-1}\int_{\mathcal{A}} Y(z)\,\mathrm{d}z$$

The minimum variance linear unbiased estimator of μ is, from de Gruijter and ter Braak (1990), the generalized least squares estimator

$$[8] \qquad \widetilde{\mu} = V_{\mathrm{m}}\left[\,\widetilde{\mu}\,\right]\mathbf{1'}\boldsymbol{C}_{\mathrm{s}}^{-1}\boldsymbol{Y}_{\mathrm{s}}(z)$$

where $\mathbf{1}$ is a unit vector of length n, $\boldsymbol{C}_{\mathrm{s}}$ is the $n \times n$ sample covariance matrix, $\boldsymbol{Y}_{\mathrm{s}}(z)$ is the n-vector of sample soil measurements, and $V_{\mathrm{m}}\left[\,\widetilde{\mu}\,\right]$ is the scalar variance of $\widetilde{\mu}$, viz.

$$V_{\mathrm{m}}\left[\,\widetilde{\mu}\,\right] = \bigl(\mathbf{1'}\boldsymbol{C}_{\mathrm{s}}^{-1}\mathbf{1}\bigr)^{-1}$$

I emphasize that in this model-based framework, soil temperature is the random variable, while sample location is regarded as fixed: even if sample locations $z_k \in \mathcal{A}$ are selected at random, that fact plays no role in estimation and inference. The upshot of this circumstance is that it may behoove the sampler to locate the z_k according to some rational design, e.g., to spread the sample out evenly over $\mathcal{A}$. In this framework the correlation among the $Y(z_k)$ variables is explicitly treated via $\boldsymbol{C}_{\mathrm{s}}$. The estimator $\widetilde{\mu}$ remains model unbiased even when the spatial correlation is ignored, say by setting $\boldsymbol{C}_{\mathrm{s}}$ to the identity matrix in eq. 8; however, $\widetilde{\mu}$ would no longer be BLU.

The best linear unbiased predictor (BLUP) of the spatial mean, $\bar{Y}$, is

$$\widetilde{Y} = \widetilde{\mu} + \bigl(\boldsymbol{C}_{\mathrm{s}}^{-1} - V_{\mathrm{m}}(\widetilde{\mu})\boldsymbol{C}_{\mathrm{s}}^{-1}\boldsymbol{J}\boldsymbol{C}_{\mathrm{s}}^{-1}\bigr)\bar{\boldsymbol{C}}_{S,A}\boldsymbol{Y}_{\mathrm{s}}(z)$$

where $\boldsymbol{J} = \mathbf{11'}$ and $\bar{\boldsymbol{C}}_{S,A}$ denotes the vector of mean covariances between each sample point and all points in $\mathcal{A}$ (de Gruijter and ter Braak 1990; also see Corsten 1989). ■

Remark 7*a*

Does $\bar{Y}$ differ from μ_y? Yes: μ_y in the design-based framework is a parameter of the fixed population whereas $\bar{Y}$ in the model-based framework is a random variable.

Remark 7*b*

It is beyond the scope of this article to discuss ways in which $\boldsymbol{C}(\boldsymbol{h})$ and hence $\boldsymbol{C}_{\mathrm{s}}$ might be parametrized; I defer to Cressie (1991), Deutsch and Journel (1992), or other texts on spatial statistics. Ripley (1981) described several spatial point processes; Diggle (1983) described tests for complete spatial randomness.

Whether μ or $\bar{Y}$ is of principal interest depends on the purpose of the investigation. The model mean may be of interest for process-oriented investigations whereas the population mean may be more relevant when interest lies in the specific region $\mathcal{A}$, not to some broader inference space (de Gruijter and ter Braak 1990). In other words, $\bar{Y}$ may be the focus of attention when a particular population within $\mathcal{A}$ is the primary concern; when the process that generated the population is the primary concern, then μ will more likely be the focus of attention. For example, if one is interested in discerning whether different combinations of harvesting and site preparation regimes affect soil temperature, then μ, not $\bar{Y}$, will be the object of interest: $\bar{Y}$ will vary among replicates of stands subjected to identical treatments.

To recap some of the salient features of model-based inference: (*i*) the population is regarded as a realization of a stochastic process, and the value y_k associated with the kth population element is a realization of a random variable, Y_k, (*ii*) the reference distribution is the probability distribution for $Y_k \in \mathcal{U}$ stipulated by the model, (*iii*) the statistical properties of estimators are deduced conditionally upon the observed sample and the stipulated model, and (*iv*) the sample design is irrelavant to inference, although probabilistic selection may help to make estimation robust to model misspecification.

One of the chief complaints lodged against using a presumed model as the basis for inference, indeed perhaps the major impediment to broader employment of model-based inference, is the potential for serious bias when the model is misspecified. This lack of robustness is the major theme explored in Hansen et al. (1983). Various tactics, such as designing the sample to be “balanced on X” (cf. Royall and

Herson 1973*a*, 1973*b*), have been proposed as a way to make inference more robust in the event of model misspecification. Nonetheless, there remains a pervasive suspicion that the model is riskier than the design as a basis for inference. In parallel to the complaint against design-based inference that the source of randomness is artificial is the following complaint voiced by Smith (1994) about the use of models:

> ...models, in the strict scientific sense allowing for replicated measurement, do not exist in much of social science. In the absense of models for the underlying social processes which are generally held to be true, model-based inferences lose all their desirable properties.

Those of us involved with devising sampling strategies to monitor recreation use, e.g., Gregoire and Buhyoff (1998), are sympathetic to this view. Moreover, many involved with multiresource inventories of biological phenomena would argue similarly to Smith — believable biological models simply are lacking in many settings and circumstances. As Olsen and Schreuder (1997, p. 174) succinctly stated, “The dynamics in forest and range lands are still much too complex for us to mimic them realistically with models, and numerous key variables cannot be measured yet in a practical manner.”

5. Line intersect sampling

Does the validity of inference for estimators of population characteristics based on line intersect sampling (LIS) depend on the random orientation of the population elements? One is certainly led to believe so from Battles et al. (1996) who stated: “A basic assumption is that the sampled objects are randomly oriented with respect to the transects.” Shiver and Borders (1996, p. 311) gave the same impression: “Note that one of the assumptions underlying line intersect sampling is that bolts lying on the ground are randomly oriented.... To avoid problems with nonrandom orientation of bolts, several transects should be run in different directions.” Many developments in the theory of LIS have indeed been predicated on the assumption that population elements, or particles in the lexicon of LIS, are both located and oriented uniformly at random in $\mathcal{A}$. However, it is hardly necessary to make such restrictive assumptions that are almost surely never satisfied in practice — Kaiser (1983) provided an elegant and thorough derivation of estimators and their properties “by assuming, as is done in sampling theory, that the population is fixed and that randomness enters the problem only through the sampling scheme.” Failure to recognize the design-based approach in LIS not only has resulted in unnecessary complications of sampling design and execution, but also in characterizations of LIS that are baseless from a design-based perspective.

In the design-based approach, randomness is introduced via the probabilistic location, and possibly orientation, of the line transects of common length L. Both the spatial arrangement and shape of the particles (bushes, rocks, grasses, etc.) in the population are regarded as fixed, along with all dimensional and size characteristics associated with them. In particular, particles need not have a convex shape.

Suppose that a transect has orientation θ with respect to some established baseline, and that its midpoint is located uniformly at random within the region $\mathcal{A}$. Kaiser (1983) showed that the conditional probability of including the kth particle, say $\mathcal{P}_k$, is

$$[9] \qquad \mathrm{Prob}(I_k = 1 \mid \theta) = \frac{w_k(\theta)L}{A}$$

where A is the area of $\mathcal{A}$ and $w_k(\theta)$ is the perpendicular distance between parallel tangents to $\mathcal{P}_k$. (If $w_k(\theta)$ is measured by calipers, the twin arms of the calipers would be oriented parallel to θ; see Fig. 3.) In eq. 9, it is assumed that the numerator and denominator are scaled to identical units of measure.

The unconditional inclusion probability of $\mathcal{P}_k$ is

$$[10] \qquad \mathrm{Prob}(I_k = 1) = E_\theta\big[w_k(\theta)\big]\,\frac{L}{A}$$

where

$$E_\theta\big[w_k(\theta)\big] = \frac{1}{\pi}\int_0^\pi w_k(\theta)\,\mathrm{d}\theta$$

and where in this context, π is the universal mathematical constant, not an inclusion probability. Let x_k be an auxiliary variable, whose value may vary among the $\mathcal{P}_k \in \mathcal{U}$ and may depend upon the orientation, θ, of the line transect. When its value depends on where $\mathcal{P}_k$ is intersected, x_k is a random variable.

Again drawing from Kaiser (1983), a conditional estimator of T_y per unit area, i.e., $\lambda = T_y/A$, is

$$\widehat{\lambda}_c = \frac{1}{A}\sum_{k\in\mathcal{U}} \frac{x_k y_k I_k}{E\big[x_k I_k \mid \theta\big]}$$

and an unconditional estimator of λ is

$$\widehat{\lambda}_u = \frac{1}{A}\sum_{k\in\mathcal{U}} \frac{x_k y_k I_k}{E_\theta\big[x_k I_k\big]}$$

Example 9

Suppose y_k is the area, a_k, of $\mathcal{P}_k$ projected onto the horizontal plane of $\mathcal{A}$. Thus, λ is ground cover expressed as a proportion, rather than a percentage. By letting x_k be the length of $\mathcal{P}_k$ coincident with the transect line:

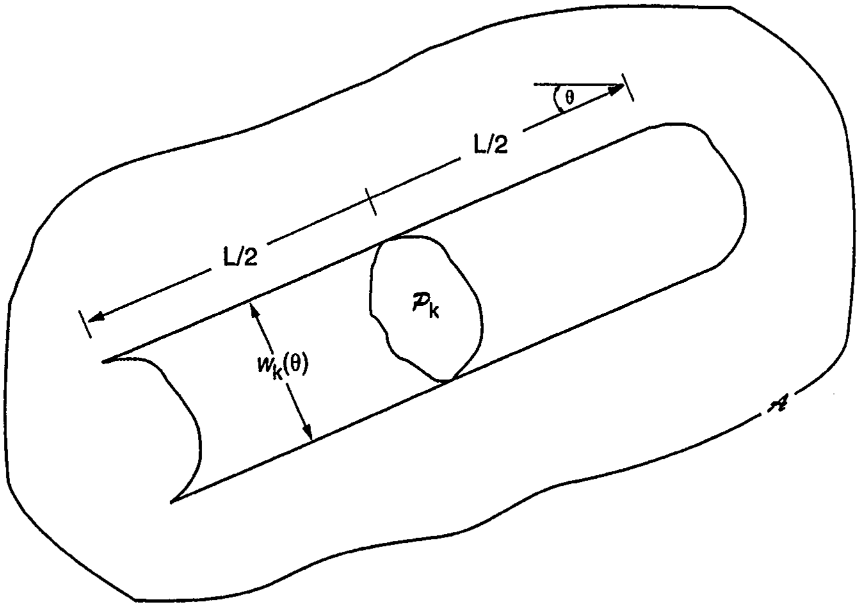
$$E\big[x_k I_k \mid \theta\big] = E_\theta\big[x_k I_k\big] = \frac{a_k L}{A}$$

Thus,

$$\widehat{\lambda}_c = \widehat{\lambda}_u = \frac{1}{L}\sum_{k\in S} x_k$$

This estimator of cover was demonstrated by Canfield (1941), although Kaiser (1983) indicated that geologists may have

Fig. 3. Particle on $\mathcal{A}$ and the maximum distance perpendicular to L that has orientation θ.



used it half a century prior to that. The judicious choice of the auxiliary variate, x_k, to be the length of intersection obviates the need for the direct and more difficult measurement of projected area, a_k, as attempted by Battles et al. (1996). ■

The conditional estimator, $\widehat{\lambda}_c$, is prudent when the sample is designed with line transects having a fixed, common orientation. If both the location and the orientation of the transects are allowed to vary randomly, then the unconditional estimator, $\widehat{\lambda}_u$, should be used. Within the design-based mode of inference, whether the transect orientation is fixed or random is a design option that need not be considered for the purpose of countering a lack of randomness of the population particles. Especially in view of foresters' longstanding familiarity with inventory of systematically planted stands of trees, a concern that trees or any other aspect of the forest be randomly placed seems rather odd, and indeed is unnecessary.

A result discussed by Mátern (1956) can be used to simplify the unconditional estimator, $\widehat{\lambda}_u$, in certain situations. Specifically, when $x_k = 1$ by design, then $E_\theta\big[x_k I_k\big]$ equates to eq. 10. Moreover, $E_\theta\big[w_k(\theta)\big] = w_k^*/\pi$, where w_k^* is the girth of the convex hull of $\mathcal{P}_k$ projected onto the horizontal plane of $\mathcal{A}$.[2] This is an exact result, providing that the actual girth of $\mathcal{P}_k$ can be measured with negligible error. As an alternative to measuring the girth of $\mathcal{P}_k$, a two-stage sampling strategy can be employed, where the methods presented by Gregoire and Valentine (1995) can be used at the second stage of sampling to provide a consistent estimate of the girth of $\mathcal{P}_k$.

Example 10

Let $x_k = y_k = 1$ for all $\mathcal{P}_k \in \mathcal{U}$. Thus, $T_y = N$, and λ is the density, N/A, of particles in $\mathcal{A}$. For a fixed orientation of the transect line:

$$\widehat{\lambda}_c = \frac{1}{L}\sum_{k\in S}\frac{1}{w_k(\theta)}$$

whereas for a random orientation

$$\widehat{\lambda}_u = \frac{\pi}{L}\sum_{k\in S}\frac{1}{w_k^*}$$

■

Example 11

Let $\mathcal{P}_k$ be the kth road within $\mathcal{A}$, and let y_k be its length. In this situation, λ is the number of kilometres of road per square kilometre. By letting x_k be the number of intersections of the line transect with $\mathcal{P}_k$, one can readily estimate λ unconditionally with

$$\widehat{\lambda}_u = \frac{\pi}{2L}\sum_{k\in S}x_k$$

This was first deduced by Mátern (1964) and later used by Skidmore and Turner (1992) to assess map accuracy. ■

Kaiser (1983) should be consulted for many other special cases and examples, all of which are derived under the traditional design-based framework of a fixed population about which no distributional assumptions are made in order to validate inference.

[2] For this reason, a diameter-tape measurement of diameter at breast height (DBH) can be regarded as the expected value of a randomly oriented calipered measurement of DBH.

Remark 9

The mirage method of Schmid-Haas (1969; also see Gregoire 1982) is well known in forest inventory as a means to avoid design-based bias that otherwise would accrue when inclusion areas of trees near the stand boundary are truncated by it. The same method has been adapted to LIS within the design-based framework by Gregoire and Monkevich (1994). Other methods of avoiding edge-effect bias exist; see Kaiser (1983) or Thompson (1992, p. 231).

In wildlife studies, a technique similar to LIS as described above is prevalent. In this context the particles are animals. Rather than being selected into the sample when the line transect actually crosses an animal, an event unlikely to occur with most live specimens, it is selected and counted when sighted by an observer traversing the transect. Estimation of animal abundance or density is based on an assumed detection function (Seber 1982, pp. 460–471). Because this assumed detection function affects the reference distribution, inference about population parameters is model based. Peilou (1985) distinguished LIS from both line intercept and line transect sampling. While line transect is used fairly consistently to indicate the sampling of wildlife populations in the manner just described, the other two terms are used interchangeably in the literature of the past few decades.

Fixed- and variable-radius plot sampling

In this penultimate section, I consider the independent placement of M sampling locations within the forested region $\mathcal{A}$. Each location serves as the center of a circular, fixed-radius plot or of a variable-radius plot, with the understanding that the same type of plot is established at all M locations. Within $\mathcal{A}$ are N trees that collectively form the population of interest. The kth tree will be denoted by $\mathcal{T}_k$, its biomass by y_k, its DBH by d_k, and the target parameter is $T_y = \sum_{k\in\mathcal{U}} y_k$.

At each location, trees are selected into the plot sample by means of a distance rule used uniformly throughout the M locations. Specifically, let Z_m denote the location of the mth plot center within $\mathcal{A}$ and let t_{km} denote the distance separating $\mathcal{T}_k$ from Z_m. Since Z_m is random, t_{km} is a random variable, also. The distance rule is this: $\mathcal{T}_k$ is selected into the sample at Z_m if $t_{km} \le v_k$, where v_k is the limiting distance of $\mathcal{T}_k$, namely

$$v_k = \begin{cases} R, \text{the plot radius, when using fixed-radius plots} \\ R_k = \phi\, d_k, \text{where } \phi \text{ is commonly known as the} \\ \text{plot radius factor (cf. Husch et al. 1982) when} \\ \text{using variable-radius plots} \end{cases}$$

Thus, if the random variable I_{km} indicates whether $\mathcal{T}_k$ is selected into the sample at location Z_m, then

$$I_{km} = \begin{cases} 1, & \text{if } t_{km} \le v_k \\ 0, & \text{otherwise} \end{cases}$$

Let $p_k = \mathrm{Prob}\big(I_{km} = 1\big)$, that evidently coincides with π_k when $m = 1$. A physical interpretation of p_k is possible which accords with the design-based paradigm that the population is fixed but the sampling location is random: p_k is the proportion of the region $\mathcal{A}$ within which a plot can be located such that $t_{km} \le v_k$. Note that p_k is a property of $\mathcal{T}_k$ fixed by the sampling design, not the sampling location Z_m, and thus, it is not a random variable but a constant. If $\mathcal{T}_k$ is closer to the edge of $\mathcal{A}$ than v_k, p_k will be less than what it would be had it been further from the edge than v_k. Plots located near the edge have absolutely no effect on p_k for all $\mathcal{T}_k \in \mathcal{U}$; see Gregoire (1982) for an elaboration of the cause of "boundary overlap," as it is commonly known.

Denote the HT estimator of T_y from the sample at Z_m by

$$\widehat{T}_{y,m} = \sum_{k\in\mathcal{U}} \frac{y_k I_{km}}{p_k}$$

which is unbiased and has variance

$$[11] \qquad V\left[\widehat{T}_{y,m}\right] = \sum_{k\in\mathcal{U}} y_k^2 \left(\frac{1-p_k}{p_k}\right) + \sum\sum_{k\neq k'\in\mathcal{U}} y_k y_{k'} \left(\frac{p_{kk'} - p_k p_{k'}}{p_k p_{k'}}\right)$$

In eq. 11, $p_{kk'}$ is the proportion of $\mathcal{A}$ within which a plot can be located such that both $t_{km} \le v_k$ and $t_{k'm} \le v_{k'}$.

Using the same estimator $\widehat{T}_y$ at each of the M locations provides M independent estimates of T_y. Moreover, $V\left[\widehat{T}_{y,m}\right] = V\left[\widehat{T}_{y,m'}\right]$ for all $m, m' = 1, \ldots, M$, as can be demonstrated by inspection of eq. 11. For convenience, let $\mathcal{V}$ represent $V\left[\widehat{T}_{y,m}\right]$, $m = 1, \ldots, M$. Finally, let

$$\widetilde{T}_y = M^{-1} \sum_{m=1}^{M} \widehat{T}_{y,m}$$

Because the $\widehat{T}_{y,m}$ are independently and identically distributed, then

$$[12] \qquad V\big[\,\widetilde{T}_y\,\big] = \mathcal{V}/M$$

While $\widetilde{T}_y$ may not be immediately recognizable, it is indeed the usual estimator of inventory. For example, when sampling with fixed-radius plots, if the area of the plot is signified by a and expressed in the same units as the area A, then except for those trees close to the edge, $p_k = a/A$, so that $\widetilde{T}_y = \big(A/aM\big)\sum_{m=1}^{M} Y_m$, where Y_m is the aggregate biomass of trees sampled on the mth plot. The variance expression eq. 11 has been corroborated by Gregoire and Scott (1990) using data from a 5.2-ha stand of 4676 trees whose location and size had been measured. These measurements enabled eq. 11 to be evaluated exactly. This value was then compared with the variance of an empirical sampling distribution comprising estimates $\widetilde{T}_y$ from 48 ,000 sample plots located uniformly at random on the 5.2-ha region via computer simulation. This variance of the empirical sampling distribution agreed with that value derived analytically in eq. 11 to within a fraction of a percent for estimators of total tree frequency, population aggregate basal area, and aggregate volume. A similar verification was undertaken by Nelson and Gregoire (1994) for two-stage sampling consisting of variable-radius plot sampling at the first stage. Here, also, the analytically derived variance matched the variance observed from the simulation to within a fraction of a percent.

I use eq. 11 to amplify a point raised in an earlier section. The spatial distribution of trees affects the joint selection probabilities $p_{kk'}$. But the variance of $\widehat{T}_{y,m}$ and hence of $\widetilde{T}_y$ does not and need not account for possible spatial correlation extant among the tree sizes or other attributes. If the analytical derivation of eq. 11 seems obscure, then at least the simulation results, cited above, should help to make this point clear.

A design-unbiased estimator of $V[\widetilde{T}_y] = \mathcal{V}/M$ is obtained as the observed variation among the $\widehat{T}_{y,m},\ m = 1, \ldots, M$:

$$\widehat{v}\big[\,\widetilde{T}_y\,\big] = \big(M(M-1)\big)^{-1} \sum_{m=1}^{M} \Big(\widehat{T}_{y,m} - \widetilde{T}_y\Big)^2$$

$$= \big(M(M-1)\big)^{-1} \left[\sum_{m=1}^{M} \widehat{T}^2_{y,m} - M\,\widetilde{T}_y^2\right]$$

The unbiasedness of $\widehat{v}\big[\,\widetilde{T}_y\,\big]$ as an estimator of $V\big[\,\widetilde{T}_y\,\big]$ is proved in the Appendix.

Eriksson (1995b) favored an infinite population approach wherein points on the ground are the sampling units. For the purpose of estimating tree characteristics, the difference between these alternate views may not lead to any important distinctions — for example, Eriksson (1995*b*) arrived at the same variance estimator as shown above using the superpopulation approach. However, for nontree characteristics, her approach may have considerable merit.

Yet another view is that of Shiver and Borders (1996) wherein the population consists of $N = A/a$ plots, from which n are selected. If the region $\mathcal{A}$ truly were to be tessellated by N plots, then this approach might be defensible. Its appropriateness to current practice is questionable, as $\mathcal{A}$ rarely comprises a set of N nonoverlapping plots of land.

Discussion

More than a decade ago in this journal, Warren (1986) exhorted scientists who publish results of statistically fitted models to state explicitly the model structure and assumptions, as well as any apparent deviation of the data from the assumed model. This paper has been written to plea for similar explicitness when reporting methods and results from survey sampling. In particular the mode of inference should be unequivocally stated and accompanied by a terse description (e.g., Gregoire and Monkevich 1994; Eriksson 1995*b*) of the reference distribution used by the authors as the basis for inference. Lacking such a statement, readers must try to gather what they can from context, and, as some of the cases cited here demonstrate, the intended or implied inferential basis is not always clear from context.

The distinction between using the sampling design versus an assumed population model has been emphasized above in order to help dispel the confusion between the two inferential paradigms. Under the design-based paradigm the reference distribution is a consequence of the sample design for a fixed population. Under the model-based paradigm the reference distribution is a consequence of a presumed model of population behavior, i.e., a superpopulation, and the sample design is ancillary to inference. Oscar Garcia (personal correspondence dated 13 April 1998) aptly reminded me that I have considered model-based inference from a frequentist perspective only; Bayes and decision-theoretic approaches are possible, also.

Is one paradigm preferable to the other? The best answer I have to offer is an equivocal one: it all depends on the circumstance and the objectives of the survey. In this regard, the following sentiment from Smith's (1994) invited address to the Washington Statistical Society is most apt. Speaking about the reconciliation of the two modes of inference, Smith asserted:

> All inferences are the product of man's imagination and there can be no absolutely correct method of inductive reasoning.... My overall conclusion is that there is no single paradigm for statistical inference and that we should concentrate on identifying the differences and enjoy the diversity of our subject.

I speculate that most of us in applied sciences such as forestry and ecology share this view and are content with it. We keep modeling tools in one apron and pull them out when needed for analytical inference, such as fitting a biomass regression equation by some optimality criterion like maximum likelihood or least squares. Another apron is kept stocked with survey tools to deduce characteristics of a specific biological population. Occasionally we get caught in a quandary when data from complex survey designs are needed to discern relationships; the purported growth decline of pines in the southern United States comes to mind in this regard. To exploit the strengths of both design-based and model-based inference, Olsen and Schreuder (1997) are trying ambitiously to combine elements of both paradigms for the purpose of establishing cause–effect relationships.

S. Stehman (personal correspondence) touched on the possible reconciliation of the two approaches when he queried whether design-based inference can be viewed as model-based inference for estimators conditioned on the event Ψ, where

$$\Psi = \big\{Y_k = y_k, k \in \mathcal{U}\big\}$$

which carries the implication that the X_k are similary fixed. I have long thought this to be true, but I say so cautiously. I have yet to consider all the implications of this viewpoint.

Some have suggested to me that an appeal to the central limit theorem to justify confidence intervals in the design-based mode indicates the assumption of a model, and therefore establishes a link between the two inferential paradigms. I do not argue vehemently against this viewpoint. I do maintain, however, that assuming an approximately Gaussian distribution (model) for the reference distribution is distinctly different from assuming distributional characteristics of the population, which in turn establishes the reference distribution. That distinction serously weakens the link in my view, as it is the distributional assumptions about the population that are the essence of model-based inference.

An important point made in various ways throughout this article is that an estimator may be biased (unbiased) in the design-based framework while unbiased (biased) in the model-based framework. Some find this result paradoxical, yet in my view, it serves usefully to emphasize that the two paradigms appeal to different reference distributions for inference. Irrespective of which inferential framework is used, the effect of measurement error can be pernicious: in my experience, measurement error rarely is random, and its presence, and the subsequent bias that it may impart, often is blithely ignored.

All but one of the many presubmission reviewers of this manuscript reported that it had helped them to appreciate the difference between design-based and model-based inference more clearly. More than one also indicated lingering uncertainty on various issues. Therefore, I close with some suggestions for further reading that may help to clarify and expand upon selected issues raised above.

While I have already cited many of the authoritative references on the basis of inference, as well as the design-based/model-based duality, there is an abundant literature not mentioned. I highly recommend the discussion sections following Basu (1978), Royall and Cumberland (1981, 1985), Hansen et al. (1983), and Brewer (1994). I have found them to be very provocative and instrumental in cementing my own views on the matter, and the last named reference provides an especially insightful account of dilemmas faced by proponents in both camps. Koch and Gillings (1982) have a brief but illuminating encyclopedia entry on statistical inference that specifically contrasts the reference distributions implicit to the two modes of inference discussed here. Kish (1987) explores the multiple meanings of the word “representative,” giving due recognition to the superb series of articles by Kruskal and Mosteller (1979*a*, 1979*b*, 1979*c*, 1980) devoted to the subject. In chapter 1, Kish eschews the model-based approach in favor of the “population bound” (design-based) approach, mainly because he associates the former with an abandonment of probabilistic selection. While the two are not necessarily linked in my view, his perspective makes for thoughtful reading. Thompson (1992) has a brief and straightforward discussion of the advantages offered by design-based and model-based approaches in various settings and for various survey objectives. Thompson (1997, chap. 5) writes at a more challenging level as she considers superpopulation inference and the related issues of likelihood, conditioning, and exchangeability.

Closely tied to the model-assisted approach to sampling and inference is the notion of asymptotic design unbiasedness (ADU). Sărndal (1980) makes an eloquent case of ADU estimators.

Sampling and estimation for continuously distributed populations have received much recent attention, perhaps as a result of the rapid development and prominence of statistical methods for the analysis of spatial data. Yet, continuous populations have been a focus of concern in the literature of systematic sampling dating back at least to Jones (1948). Bartlett (1986) considers a superpopulation approach to estimating the population total of a continuous population, and he makes an illuminating connection to the geostatistical technique of kriging. McArthur (1987) undertakes a simulation study of various sampling strategies to estimate the average level of response surface of a pollutant; he recommendes both stratified random sampling and importance sampling as the most precise of those he considered. Cordy (1993) providedsthe theory for HT estimation of parameters of a fixed, continuous population based on spatial probability sampling. An important paper both for its readability and its breadth of coverage of issues relating to spatial sampling of the environment is that of Cox et al. (1997).

Acknowledgements

The author is indebted to Bob Bailey, John J. Battles, Mary Christman, Ray Czaplewski, Oscar Garcia, Jeff Gove, Juha Lappi, Steen Magnussen, Tim Max, Ron McRoberts, Richard G. Oderwald, Jeff Pontius, Andrew Robinson, Hans Schreuder, Chip Scott, Steve Stehman, and Harry Valentine for extensive and insightful comments that have helped improve the quality of this paper considerably. While he dearly would like to assign responsibility for any errors remaining in the manuscript to this cohort, editorial prudence dictates that culpability be ascribed to himself alone.

References

Bartlett, R.F. 1986. Estimating the total of a continuous population. J. Stat. Planning Inference, **13**: 51–66.

Basu, D. 1978. Relevance of randomness in data analysis. *In* Survey sampling and measurement. *Edited by* N.K. Namboodiri. Academic Press, New York. pp. 267–292.

Battles, J.J., Dushoff, J.G., and Fahey, T.J. 1996. Line intersect sampling of forest canopy gaps. For. Sci. **42**: 131–138.

Bebbington, A.C. 1975. A simple method of drawing a sample without replacement. Appl. Stat. **24**: 136.

Bellehumeur, C., Legendre, P., and Marcotte, D. 1997. Variance and spatial scales in a tropical rain forest: changing the size of the sampling units. Plant Ecol. **130**: 89–98.

Bickel, P.J., and Doksum, K.A. 1977. Mathematical statistics: basic ideas and selected topics. Holden-Day, Oakland, Calif.

Box, G.E.P., and Anderson, S.L. 1955. Permutation theory in the derivation of robust criteria and the study of departures from assumptions. J. R. Stat. Soc. Ser. B (Methodol.), **17**: 1–26.

Brewer, K.R.W. 1963. Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. Aust. J. Stat. **5**: 93–105.

Brewer, K.R.W. 1994. Survey sampling inference: some past perspectives and present prospects. Pak. J. Stat. **10**: 213–233.

Brus, D.J., and de Gruijter, J.J. 1993. Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. Environmetrics, **4**: 123–152.

Canfield, R.H. 1941. Application of the line interception method in sampling range vegetation. J. For. **39**: 388–394.

Cassel, C.-M., Särndal, C.-E., and Wretman, J.H. 1977 Foundations of inference in survey sampling. Wiley, New York.

Cochran, W.G. 1977. Sampling techniques. Wiley, New York.

Cordy, C.B. 1993. An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. Stat. Prob. Lett. **18**: 353–362.

Corsten, L.C.A. 1989. Interpolation and optimal linear prediction. Stat. Neerlandica, **43**: 69–84.

Cox, D.D., Cox, L.H., and Ensor, K.B. 1997. Spatial sampling and the environment: some issues and directions. Environ. Ecol. Stat. **4**: 219–233.

Cressie, N.A.C. 1991. Statistics for spatial data. Wiley, New York.

de Gruijter, J.J., and ter Braak, C.J.F. 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Math. Geol. **22**: 407–415.

Deutsch, C.V., and Journel, A.G. 1992. GSLIB geostatistical software library and user's guide. Oxford University Press, New York.

Diggle, P.J. 1983. Statistical analysis of spatial point patterns. Academic Press, London.

Eriksson, M. 1995*a*. Compatible and time-additive change component estimators for horizontal-point sampling. For. Sci. **41**: 796–822.

Eriksson, M. 1995*b*. Design-based approaches to horizontal-point sampling. For. Sci. **41**: 890–907.

Fisher, R.A. 1935. The design of experiments. Oliver Boyd, Edinburgh.

Fisher, R.A. 1956. Statistical methods and scientific inference. Hafner Publishing Co., New York.

Fisher, R.A. 1973. Statistical methods and scientific inference 3rd ed. Hafner Press, New York.

Godambe, V.P. 1955. A unified theory of sampling from finite populations. J. R. Stat. Soc. Ser. B (Methodol.), **17**: 269–278.

Gregoire, T.G. 1982. The unbiasedness of the mirage correction procedure for boundary overlap. For. Sci. **28**: 504–508.

Gregoire, T.G., and Buhyoff, G.J. 1998. Sampling and estimating recreation use. U.S. For. Serv. Gen. Tech. Rep. PNW. In press.

Gregoire, T.G., and Monkevich, N.S. 1994. The reflection method of line intercept sampling to eliminate boundary bias. Environ. Ecol. Stat. **1**: 219–226.

Gregoire, T.G., and Schabenberger, O. 1998. Sampling skewed populations: failure rates of confidence intervals for the population total. Ecology. In press.

Gregoire, T.G., and Scott, C.T. 1990. Sampling at the stand boundary: a comparison of the statistical performance among eight methods. *In* Research in forest inventory, monitoring, growth and yield. *Edited by* H.E. Burkhart, G.M. Bonnor, and J.J. Lowe. Publ. No. FWS-3-90, School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, Va.

Gregoire, T.G., and Valentine, H.T. 1995. A sampling strategy to estimate the area and perimeter of irregularly shaped planar regions. For. Sci. **41**: 470–476.

Gregoire, T.G., Valentine, H.T., and Furnival, G.M. 1993. Estimation of bole surface area and bark volume with Monte Carlo methods. Biometrics, **49**: 653–660.

Gregoire, T.G., Valentine, H.T., and Furnival, G.M. 1995. Sampling methods to estimate foliage and other characteristics of individual trees. Ecology, **76**: 1181–1194.

Hájek, J. 1960. Limiting distributions in simple random sampling from a finite populations. Publ. Math. Inst. Hung. Acad. Sci. **5**: 361–374.

Hájek, J. 1981. Sampling from a finite population. Marcel Dekker, New York.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. 1953. Sample survey methods and theory. Vol. I: methods and applications. Wiley, New York.

Hansen, M.H., Madow, W.G., and Tepping, B.J. 1983. An evaluation of model-dependent and probability sampling inferences in survey sampling. J. Am. Stat. Assoc. **78**: 776–793.

Husch, B., Miller, C.I., and Beers, T.W. 1982 Forest mensuration. 3rd ed. John Wiley & Sons, New York.

Jones, A.E. 1948. Systematic sampling of continuous sampling populations. Biometrika, **35**: 283–290.

Kaiser, L. 1983. Unbiased estimation in line-intercept sampling. Biometrics, **39**: 965–976.

Kangas, A. 1994. Classical and model based estimators for forest inventory. Silva Fenn. **28**: 3–14.

Kish, L. 1987. Statistical design for research. John Wiley & Sons, New York.

Koch, G.C., and Gillings, D.B. 1982. Line intercept sampling; line intersect sampling; line transect sampling. *In* Encyclopedia of statistical sciences. Vol. 4. *Edited by* S. Kotz and N.L. Johnson. John Wiley & Sons, New York.

Kruskal, W., and Mosteller, F. 1979*a*. Representative sampling. I: Non-scientific literature. Int. Stat. Rev. **47**: 13–24.

Kruskal, W., and Mosteller, F. 1979*b*. Representative sampling. II: Scientific literature, excluding statistics. Int.

Stat. Rev. **47**: 111–127.

Kruskal, W., and Mosteller, F. 1979*c*. Representative sampling. III: The current statistical literature. Int. Stat. Rev. **47**: 245–265.

Kruskal, W., and Mosteller, F. 1980. Representative sampling. IV: The history of the concept in statistics, 1895–1939. Int. Stat. Rev. **48**: 169–195.

Lanke, J. 1975. Some contributions to the theory of survey sampling. AV-Centralen i Lund.

Mandallaz, D. 1991. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. Chair of Forest Inventory and Planning, Swiss Federal Institute of Technology (ETH), Zurich.

Mátern, B. 1956. On the geometry of the cross-section of a stem. Medd. Statens Skogsforskningsinst. **46**: 11.

Mátern, B. 1960. Spatial variation. Medd. Statens Skogsforskningsinst. Band 49, No. 5. [A corrected version was reprinted in 1981 under the same title by Springer-Verlag as Lect. Notes Stat. 36.]

Mátern, B. 1964. A method of estimating the total length of roads by means of a line survey. Stud. For. **18**: 68–70.

McArthur, R.D. 1987. An evaluation of sample designs for estimating a locally concentrated pollutant. Commun. Stat.: Simul. **16**: 735–759.

Nelson, R., and Gregoire, T.G. 1994. Two-stage forest sampling: a comparison of three procedures to estimate aggregate volume. For. Sci. **40**: 247–266.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. R. Stat. Soc. **97**: 558–606.

Olsen, A.R., and Schreuder, H.T. 1997. Perspectives on large-scale natural resource surveys when cause–effect is a potential issue. Environ. Ecol. Stat. **4**: 167–180.

Overton, S., and Stehman, S.V. 1995. The Horwitz–Thompson theorem in a unified perspective for probability sampling: with examples from natural resource sampling. Am. Stat. **49**: 261–268.

Peilou, E.C. 1985. Line intercept sampling; line intersect sampling; line transect sampling. *In* Encyclopedia of statistical sciences. Vol. 5. *Edited by* S. Kotz and N.L. Johnson. John Wiley & Sons, New York.

Rao, J.N.K. 1985. Conditional inference in survey sampling. Surv. Methodol. **11**: 15–31.

Rao, J.N.K. 1997. Developments in sample survey theory: an appraisal. Can. J. Stat. **25**: 1–21.

Rennolls, K. 1981 The total area of woodland in Berkshire is. . . Statistician, **30**: 275–287.

Rennolls, K. 1982. The use of superpopulation–prediction methods in survey analysis, with application to the British national census of woodlands and trees. *In* In place resource inventories: principles and practices. *Edited by* T.B. Brann, L.O. House IV, and H.G. Lund. 9–14 Aug. 1981, Orono, Me. Society of American Foresters, Bethesda, Md. pp. 395–401.

Ripley, B.D. 1981. Spatial statistics. Wiley, New York.

Royall, R.M. 1970. On finite population sampling under certain linear regression models. Biometrika, **57**: 377–387.

Royall, R.M. 1971. Linear regression models in finite population sampling theory. *In* Foundations of statstical inference. Holt, Rinehart and Winston of Canada, Toronto, Ont.

Royall, R.M., and Cumberland, W.G. 1981. An empirical study of the ratio estimator and estimators of its variance. J. Am. Stat. Assoc. **76**: 66–88.

Royall, R.M., and Cumberland, W.G. 1985. Conditional coverage properties of finite population confidence intervals. J. Am. Stat. Assoc. **80**: 355–359.

Royall, R.M., and Herson, H.J. 1973*a*. Robust estimation in finite populations. I. J. Am. Stat. Assoc. **68**: 880–889.

Royall, R.M., and Herson, H.J. 1973*b*. Robust estimation in finite populations. II. J. Am. Stat. Assoc. **68**: 890–893.

Särndal, C.–E. 1978. Design-based and model-based inference in survey sampling. Scand. J. Stat. **5**: 27–52.

Särndal, C.–E. 1980. On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika, **67**: 639–650.

Särndal, C.–E., Swensson, B., and Wretman, J. 1992. Model assisted survey sampling. Springer-Verlag, New York.

Schmid-Haas, P. 1969. Stichproben am waldrand [Sampling at the edge of the forest]. Schweiz. Aust. Forstl. Versuchswes. Mitt. **45**: 234–303.

Schreuder, H.T., and Williams, M. 1995. Design-based estimation of forest volume within a model-based sample selection framework. Can. J. For. Res. **25**: 121–127.

Schreuder, H.T., Gregoire, T.G., and Wood, G.B. 1993. Sampling methods for multiresource forest inventory. Wiley, New York.

Seber, G.A.F. 1982. The estimation of animal abundance. 2nd ed. Charles Grifffin & Company, Ltd., London.

Shiver, B.D., and Borders, B.E. 1996. Sampling techniques for forest resource inventory. Wiley, New York.

Skidmore, A.K., and Turner, B.J. 1992. Map accuracy assessment using line intersect assessment. Photogramm. Eng. Remote Sens. **58**: 1453–1457.

Smith, T.M.F. 1991. Post-stratification. Statistician, **40**: 315–323.

Smith, T.M.F. 1994. Sample surveys 1975–1990; an age of reconciliation. Int. Stat. Rev. **62**: 5–34.

Stevens, D.L., Jr. 1997. Variable density grid-based sampling designs for continuous spatial populations. Environmetrics, **8**: 167–195.

Stuart, A. 1976. Basic ideas of scientific sampling. 2nd ed. Griffin's statistical monographs and courses No. 4. Hafner Press, New York.

Sukhatme, P.V., and Sukhatme, B.V. 1970. Sampling theory of surveys with applications. Iowa State University Press, Ames, Iowa.

Thompson, M.E. 1997. Sampling. John Wiley & Sons, New York.

Thompson, S.K. 1992. Theory of sample surveys. Chapman & Hall, London.

Thomson, I. 1978. Discussion of C-E Särndal's paper. Scand. J. Stat. **5**, 43–45.

Warren, W.G. 1986. On the presentation of statistical analysis: reason or ritual. Can. J. For. Res. **5**, 43–45.

Wood, G.B., Schreuder, H.T., and Brink, G.E. 1995. Comparison of a model-based sampling strategy with point-Poisson sampling on a timber management area in the Arapahoe–Roosevelt National Forest in Colorado. Can. J. For. Res. **15**: 83–86.

Appendix

Using $\mathcal{V}$ to represent $V\left[\widehat{T}_{y,m}\right]$ for $m = 1, \ldots, M$, the expected value of $\widehat{v}\left[\,\widetilde{T}_y\,\right]$ is deduced as

$$E\left[\widehat{v}\left[\,\widetilde{T}_y\,\right]\right] = \big(M(M-1)\big)^{-1}\left[E\left(\sum_{m=1}^{M}\widehat{T}^2_{y,m} - M\big(\widetilde{T}_y^2\big)\right)\right] = \big(M(M-1)\big)^{-1}\left[\sum_{m=1}^{M}E\left(\widehat{T}^2_{y,m}\right) - M\left(E\left[\widetilde{T}_y^2\right]\right)\right]$$

$$= \big(M(M-1)\big)^{-1}\left[\sum_{m=1}^{M}\left(V\big[\widehat{T}_{y,m}\big] + T_y^2\right) - M\left(V\big[\widetilde{T}_y\big] + T_y^2\right)\right]$$

$$= \big(M(M-1)\big)^{-1}\left[M\mathcal{V} - MV\big[\widetilde{T}_y\big]\right] = \big(M(M-1)\big)^{-1}\left[M^2V\big[\widetilde{T}_y\big] - MV\big[\widetilde{T}_y\big]\right]$$

$$= \big(M(M-1)\big)^{-1}\left[\big(M(M-1)\big)V\big[\widetilde{T}_y\big]\right]$$

$$= V\big[\widetilde{T}_y\big]$$